Education
Quality and
Accountability
Office

The October 2002 Ontario Secondary School Literacy Test: Technical Explanation of the Equating Models and the Process for Determining the Passing Score

The October 2002 Ontario Secondary School Literacy Test: Technical Explanation of the Equating Models and the Process for Determining the Passing Score

Introduction

This technical paper is a companion piece to the Education Quality and Accountability Office's (EQAO's) *October 2002 Ontario Secondary School Literacy Test: Summary of the Equating Models and the Process for Determining the Passing Score*. In addition to the information provided in the summary paper, this paper explains the technical processes used to link the October 2002 test (e.g., test blueprint, scoring keys and guides, passing score) and the February 2002 test. This information is provided in two parts. Part I describes the reading component. Part II deals with the writing component.

Rationale for Equating

Test equating is required when results from one test administration need to be comparable or used interchangeably with the results from another test administration. Equating adjusts for variations in difficulty among tests that are designed to be similar in content coverage and difficulty (Kolen & Brennan, 1999; Yang & Houang, 1996).

Part I: Reading Component

Test Blueprint

The test items for the reading component are derived from the larger curriculum subject areas and expectations covered in *The Ontario Curriculum* across all subjects up to the end of Grade 9. The recently published Curriculum Connections (2003) show samples of the connections between the content of the Ontario Secondary School Literacy Test (OSSLT) and *The Ontario Curriculum*.

The test blueprint for the reading component can be described in terms of the three types of reading selections and the three skill types, as shown in Table 1 below.

Table 1: Blueprint for Reading Component by Skill Type

	English-Language Reading Selection Types				French-Language Reading Selection Types			
Skill Type	**Information**	**Graphic**	**Narrative**	**Item Total**	**Information**	**Graphic**	**Narrative**	**Item Total**
Skill 1	15	8	7	30	15	8	7	30
Skill 2	20	11	14	45	20	11	14	45
Skill 3	10	6	9	25	10	6	9	25
Item Total	45	25	30	100	45	25	30	100

Note: Skill 1 = understands directly stated ideas and information (explicit); Skill 2 = understands indirectly stated ideas and information (implicit); and Skill 3 = makes connections between personal experiences and the ideas and information in the reading selections (interpretation and extension).

There are 40 multiple-choice questions, 35 short-answer questions and 25 questions requiring an explanation. The multiple-choice and short-answer questions were scored as follows: 0 points for an incorrect response and 2 points for a correct response. The questions requiring an explanation were scored on a three-point scale: 0 points for an incorrect response, 1 point for a partially correct response and 2 points for a correct response.

Test Development

Experienced educators knowledgeable of *The Ontario Curriculum* developed the reading items. All reading items used for the October 2002 test were first pilot tested in March 2000 in schools drawn from various parts of the province that included schools with both the academic and applied programs. (Schools were asked to participate on a voluntary basis.) The items were then revised, where necessary, and further field tested in the February 2002 test, using a spiral design; that is, during the February 2002 test administration, students across the province were systematically and randomly assigned different field-test materials embedded in the live test. This approach to field testing is similar to the model used in many other jurisdictions (Smisko, Twing, & Denny, 2000). The spiral design allowed EQAO to field test adequate assessment materials for the October 2002 test, including the required backup materials. After the field test, the items were analyzed and selected based on the test blueprint as well as a number of statistical requirements (e.g., Crocker & Algina, 1986; see *XCALIBRE User's Manual*).

Psychometric Requirements ("Guidelines") for Equating

There is a plethora of literature on the psychometric requirements for equating assessments (e.g., Crocker & Algina, 1986; Kolen & Brennan, 1999; Muraki, Himbo & Lee, 2000; Yang & Houang, 1996). According to Crocker and Algina, there are three main criteria for choosing an equating procedure: "tenability of assumptions, practicality, and accuracy" (p. 464). However, Kolen and Brennan (1999, pp. 283–285) provide a more comprehensive list of conditions ("guidelines") conducive to satisfactory equating. Table 2 provides an abridged version of Kolen and Brennan's list of conditions and how they can be used to contextualize the equating process for the reading component of the October 2002 test.

Table 2: Psychometric Requirements ("Guidelines") for Equating

Requirements	Description of Requirements (source: Kolen & Brennan, 1999)	OSSLT, October 2002, Reading Context
Goals of equating	Should have clearly specified goals for equating, especially for equating accuracy and the extent to which scores are to be comparable over long periods of time.	EQAO's long-term goal is to develop and administer assessment instruments that are consistent with the expectations of *The Ontario Curriculum*. Items for each assessment instrument are developed according to the test blueprint.
Test blueprint	Test content and statistical specifications should be well-defined and stable over time.	The test content and statistical specifications are the same in the February 2002 and October 2002 administrations. The same classical and item response theory parameters were used to guide item selection and revision.
Test length	Test should be reasonably long (at least 30 items).	The OSSLT reading component consists of 100 reading items (40 multiple-choice, 35 short-answer and 25 short-answer with explanation).
Scoring keys	Scoring keys should be stable when items are used on multiple occasions.	Scoring keys and guides were piloted in March 2000 and validated during the marking of the field-test materials in February/March 2002.

Requirements	Description of Requirements (source: Kolen & Brennan, 1999)	OSSLT, October 2002, Reading Context
Content representation	The content and statistical characteristics of the common item set should be representative of the total test.	Items in the February 2002 and October 2002 tests are from similar subject areas and expectations covered in *The Ontario Curriculum*.
Proportion of common items	The common item set should be of sufficient length, at least 20% of the test for tests of 40 items or more.	All October 2002 reading passages were field tested together with the February 2002 full assessment, using a spiral design. Each of the reading passages used in the October 2002 live test has a counterpart in the February 2002 live test. Commonality was operationalized in terms of content congruency and selection types.
Placement	The common item set should be in approximately the same position in the old and new forms.	Because part of the OSSLT is released, this requires replacement with other selections. This creates a practical limitation. Furthermore, field testing a complete test would require additional school time and increased test costs (e.g., printing, marking).
Timing	Double linking: One old form should be administered during the same time of year as the form to be equated. One old form should be administered within the last year.	The regular administration of the OSSLT to Grade 10 students is in October (first semester); the 2001–2002 school year was an exception because of a security breach that occurred in October 2001. Field testing took place in February 2002 — eight months prior to the October 2002 test.
Test administration	Test and test items should be secure. Test should be administered under carefully controlled, standardized conditions that are the same each time the test is administered.	The OSSLT is administered according to requirements specified in the *Guide for Administering the Ontario Secondary School Literacy Test, October 2001* and the *Guide for Administering the Ontario Secondary School Literacy Test, October 2002*.

In assessing the context of the equating models used for the reading component of the October 2002 test, two key questions to consider are a) Must all of these conditions be met? and b) To what extent do the development and construction of the reading component meet or not meet these guidelines?

Question a): Must all these conditions be met? According to Kolen and Brennan (1999), "satisfactory equating does not require that all these conditions hold" (p. 283). This is a perspective shared by Traub (1994), as shown in the following excerpt:

> "In many applications, however, it may not be important even to try to get a yes or no answer to the question of parallelism for a set of tests. Instead, we may decide, after considering the available evidence — test means, test standard deviations, and test covariances — that two or more instruments are close enough to being parallel to reap the benefits of assuming they are really parallel...." (p. 54).

Question b): To what extent do the development and construction of the reading component meet or not meet the guidelines provided by Kolen and Brennan? It is clear that the February 2002 and the October 2002 tests are equivalent and the scores generated by these two tests are comparable, for the following reasons:

1) both the February 2002 test and the October 2002 test were developed according to the same test blueprint, which is based on *The Ontario Curriculum* (test content congruency);
2) the February 2002 test and the October 2002 test were both first piloted in March 2000, using the same student cohort drawn from various parts of the province;

3) the scoring keys and guides for the October 2002 test were first developed in March 2000 and validated in February 2002 together with the February 2002 scoring keys and guides;
4) all the reading selections used in the October 2002 test were embedded in the February 2002 test, the reading items were analyzed together with the February 2002 full-assessment reading items and the same statistical requirements were used to select these items;
5) the items from the February 2002 test and the October 2002 test were concurrently calibrated using item response theory analysis and
6) the test administration for both tests was conducted using standardized procedures and under the same secure conditions.

Equating Models for the Reading Component

The equating methods used to derive equivalent passing scores for the October 2002 test and the February 2002 test can be grouped into three categories: 1) classical test theory (CTT) analysis, 2) Item Response Theory (IRT) analysis and 3) cluster analysis. Each of these methods is described below.

1) Classical Test Theory Analysis

Three techniques were used: linear equating, the equi-percentile method and mean equating.

Linear Equating. Scores on two tests are considered equated if they have the same number of standard deviations above or below the mean in a group of examinees. This process is well documented in the assessment literature (e.g., Crocker & Algina, 1986; Kolen & Brennan, 1999; Petersen, Kolen & Hoover, 1989). For instance, see the excerpt below on the definition, formula and computation process:

> **Definition:** "In linear equating, a transformation is chosen such that scores on two tests are considered to be equated if they are correspond to the same number of standard deviations above (or below) the mean in some group of examinees" (Angoff [1994] quoted in Petersen, Kolen & Hoover, 1989, p. 251).
>
> **Formula:** "$(x - \mu_x)/\delta_x = (y - \mu_y)/\delta_y$.
> Where, x and y refer to scores on the test forms to be equated and μ and δ refer to means and standard deviations, respectively, of x and y in some population of examinees" (p. 251).
>
> **Computation process:**
> a) Compute the sample means (m) and sample standard deviations (s) of scores on the two forms to be equated
> b) Obtain equated scores on the two forms by substituting these values in the above equation (p. 252)

Equi-Percentile Method. This method determines which student populations have the same percentile rank on the two tests. A detailed discussion of this technique is provided by Kolen and Brennan (1999). See the excerpt below on how this technique is operationalized.

"An equi-percentile equivalent for a population of examinees can be constructed in the following manner: For a given Form X score, find the percentage of examinees earning scores at or below that Form X score. Next, find the Form Y score that has the same percentage of examinees at or below it. These Form X and Form Y scores are considered to be equivalent" (Kolen & Brennan, 1999, pp. 36–37).

Mean Equating. The scores on the second test are converted so that the mean is equal to the mean of the first test; the difference between the means is set as a constant across the score scale (Barnard, 1996).

2) Item Response Theory Analysis

The items from the February 2002 and October 2002 tests were calibrated together using responses from the February administration. Estimates of student ability were derived for the February examinees. Using the ability estimates and students' results, it was determined how the same students would score if they wrote the October 2002 test. In this way, students' IRT scores for the February 2002 and October 2002 tests were placed on the same scale and were made directly comparable.

A description of the steps involved in the IRT equating process follows:

Step 1: IRT Item Calibration

Two scoring approaches were used: one for items scored dichotomously (i.e., 0 = incorrect, 2 = correct) and another for items scored polytomously using three categories (i.e., 0 = incorrect, 1 = partially correct, 2 = correct). The dichotomously scored items were calibrated using a one parameter logistic model that sets the guessing parameter at 0 and constrains the discrimination parameter to be equal for all items. The polytomously scored items were calibrated using the generalized partial credit model (Bock, 1972).

Step 2: Linear Transformation of the Ability (Base IRT) Scale

Since the October 2002 test was formed from items that had been field tested together with the February 2002 test, the items from the two tests were calibrated using examinee responses from the February 2002 administration. Student ability estimates were derived as part of this item calibration. The ability estimate for the February 2002 examinees was then used to estimate the score one would expect them to have achieved if they had been given the October 2002 test. This joint-item calibration process eliminates the need for linear transformation of the ability scale.

Step 3: Converting October 2002 Scores to February 2002 Scores

After obtaining the item parameters for all students in each language group (i.e., one for English-language and another for French-language), each student's achievement on the February 2002 test was scored. Student IRT scores thus reflected the number of items a student answered, the correctness of his or her responses and the characteristics of the items.

The two test characteristic curves (TCCs), one for the February 2002 test and another for the October 2002 test, were "mapped" onto the same ability scale and converted back to number-correct scoring.

Figures 1 and 2 show the TCCs for English-language and French-language, respectively.

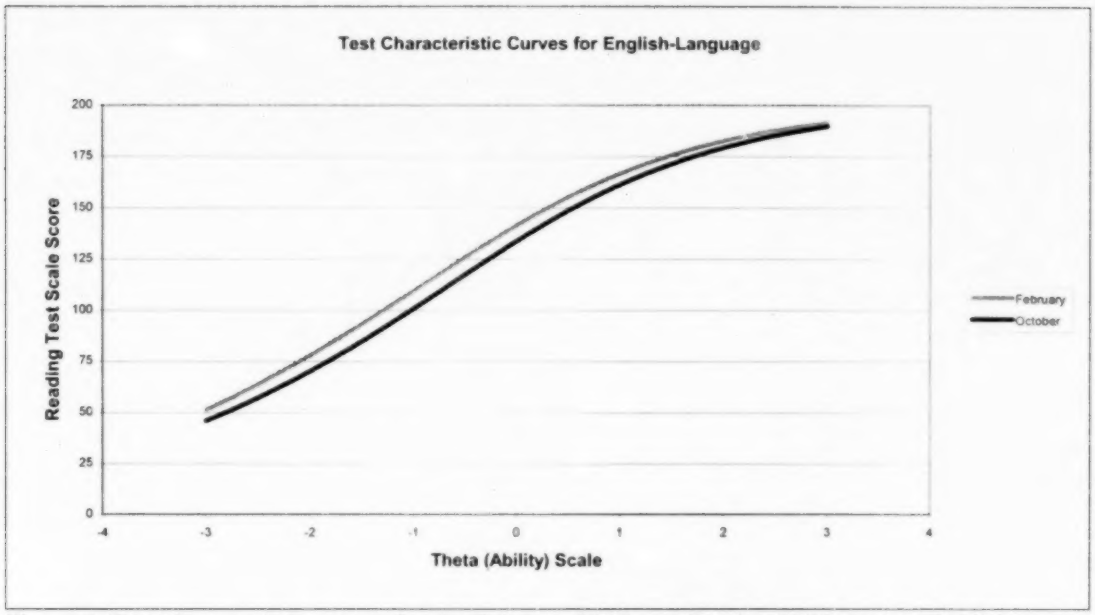

Figure 1. Test characteristic curves for the February 2002 and October 2002 tests — English-Language.

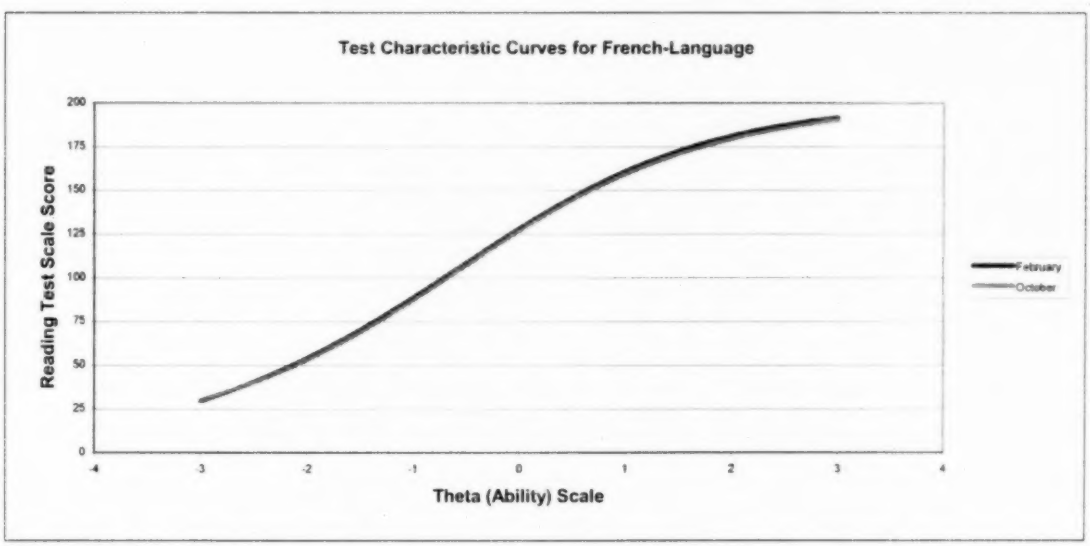

Figure 2. Test characteristic curves for the February 2002 and October 2002 tests — French-Language.

3) Cluster Analysis

Cluster analysis is both an exploratory and confirmatory technique. In an exploratory analysis, cluster analysis can be used when passing score points are not known. As a confirmatory analysis, this technique can be used when passing scores are known and need to be validated (Sireci, 2001).

The excerpt below provides a definition of cluster analysis and its application to test score analysis.

> "[The object of cluster analysis] is to sort cases (people, things, events, etc.) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong." (Source: http://www.clustan.com/what_is_cluster_analysis.html)
>
> "If test takers are considered as 'objects' to be clustered, the goal of cluster analysis is to group them into categories such that test takers within each category are similar to one another, but different from test takers in other categories" (Sireci, 2001, p. 341).

For the October 2002 test, cluster analysis was based on the three question type sub-scores (i.e., multiple-choice, short-answer and short-answer with explanation scores). Along the performance continuum, a passing score was derived to separate successful and unsuccessful students.

Final Reading Passing Score Results

Using the CTT analysis and the cluster analysis techniques, it was evident that a passing score range could be between 117 and 120 for the English-language students and between 118 and 124 for the French-language students. The IRT analysis was used to determine the final passing score: 119 out of a maximum of 200 points for the English-language students and 118 out of a maximum of 200 points for the French-language students. These passing scores are equivalent to the passing scores from the February 2002 test. Although there are pros and cons to the various equating methods (Barnard, 1996, Kolen & Brennan, 1999), the advantage of using IRT analysis is that it allows for the concurrent calibration of items from different forms. The two sets of scores are placed on the same scale, allowing results to be directly comparable.

Part II: Writing Component

The writing tasks are derived from the larger curriculum subject areas and expectations covered in *The Ontario Curriculum* across all subjects up to the end of Grade 9. The recently published Curriculum Connections (2003) show samples of the connections between the OSSLT content and *The Ontario Curriculum*.

As with the reading component of the OSSLT, Kolen and Brennan's list of conditions for equating can be applied to the writing component to contextualize the comparability between the February 2002 test and the October 2002 test. Two key questions to consider are a) How do the writing tasks on the February 2002 and the October 2002 tests compare? and b) Can the writing scores derived from the two tests be equated?

Table 3: Psychometric Requirements ("Guidelines") for Equating

Requirements	Description of Requirements (source: Kolen & Brennan, 1999)	OSSLT, October 2002, Writing Context
Goals of equating	Should have clearly specified goals for equating, especially for equating accuracy and the extent to which scores are to be comparable over long periods of time.	EQAO's long-term goal is to develop and administer assessment instruments that are consistent with the expectations of *The Ontario Curriculum*. Tasks for each assessment instrument are developed according to the test blueprint.
Test blueprint	Test content should be well-defined and stable over time.	For the February 2002 and the October 2002 tests, four tasks of the same genre comprise the writing component: • summary • series of paragraphs expressing an opinion • news report • information paragraph
Test length	Test length should be standardized.	No change between the February 2002 test and the October 2002 test: • summary — fewer than 100 words • opinion piece — minimum of three paragraphs; three lined pages provided, indicating approximate length required • news report — one lined page provided, indicating approximate length required • information paragraph — lined space provided, approximately one page in length
Scoring keys	Scoring rubrics should be stable.	No change in scoring rubrics used for holistic marking between the February 2002 test and the October 2002 test.
Content representation	The content characteristics of the writing tasks should be representative of the writing component of the test.	Content representation criteria comparable between the February 2002 test and the October 2002 test. – Canadian content – student- and/or school-centered topics – audience: adults for three tasks (summary, opinion piece, information paragraph) – information paragraph: student response to include at least four points as supporting details from 18 points provided
Placement	The writing tasks should be in the same position in the old and new forms.	The writing tasks were placed in the same position in the October 2002 as in the February 2002 test.
Timing	Double linking: One old form should be administered during the same time of year as the form to be equated. One old form should be administered within the last year.	The regular administration of the OSSLT to Grade 10 students is in October (first semester); the 2001–2002 school year was an exception because of a security breach that occurred in October 2001. Field testing took place in February 2002 — eight months prior to the October 2002 test.
Test administration	Test should be secure. Test should be administered under carefully controlled, standardized conditions that are the same each time the test is administered.	The OSSLT is administered according to requirements specified in the *Guide for Administering the Ontario Secondary School Literacy Test, October 2001* and the *Guide for Administering the Ontario Secondary School Literacy Test, October 2002*. No change in the administration process with respect to the number of days or the sequence of booklets and tasks. Test is two half-days for each administration.

Passing-Score Determination and Results for Writing

The standard-setting methodology known as the booklet classification was used to derive equivalent passing scores for the October 2002 and the February 2002 tests. This process

involved working with successful and unsuccessful student work from the February 2002 assessment and equating this material to marked material from the October 2002 assessment. The equating process took place in November 2002. The standard-setting methodology is described later. Below is the marking process used for the October 2002 test.

The Writing Marking Process. There were two phases to marking the writing component.

> *Holistic Phase 1 Marking.* Each student's work was marked holistically, meaning that as each task was marked, all characteristics of the task were considered in combination and scored. A different marker marked each task.
>
> Markers were trained on scoring rubrics unique to each task. Each rubric contained descriptors of student work along the scoring continuum. Examples of student work representing each point along the continuum were used in conjunction with the rubrics. For each of the four writing tasks (summary, series of paragraphs expressing an opinion, news report, information paragraph) a score of 0, 1, 2, 3 or 4 was assigned using a holistic rubric.
>
> *Holistic Phase 2 Marking.* Student work that did not meet the passing criteria in Phase 1 was marked again holistically. If the work received a passing mark in Phase 2, the student passed the writing component.
>
> For each phase, the holistic scoring process involved four different markers scoring a student's work, yielding four holistic scores, one for each of the four writing tasks. The overall holistic score was obtained by summing the individual scores for the tasks.
>
> *Analytic Marking.* Student papers that were still not successful in Phase 2 holistic marking were further marked using an analytic marking rubric. The purpose of this analytic marking was to give detailed feedback to the student on his or her work for remediation purposes.

Writing Standard-Setting Process:
Linking the February 2002 and the October 2002 Tests

The standard-setting exercise, which used a booklet-classification methodology, took place on November 19–20, 2002, at the EQAO marking site. Eleven panelists were involved in the activity — six English-language experts and five French-language experts. The eleven experts came from diverse backgrounds: EQAO subject experts, university personnel, private sector representatives and Ministry of Education personnel.

A description of the steps in the standard-setting procedure follows:

Step 1: Training the Panelists

The panelists were trained by a team composed of three EQAO education officers who were involved in developing the test and rubrics and a psychometrician from the University of Toronto. The training included a thorough familiarization with the test content and expectations,

the rubrics for the February 2002 and October 2002 tests and samples of students' work exemplifying a wide range of writing performances. There was also some practice in rating student work using the scoring rubrics.

Step 2: Forming Groups
The panelists were grouped into four teams — two English teams and two French teams. The English teams worked on English writing papers while the French teams worked on French writing papers. Interaction between the teams was minimized.

Step 3: Examination of the February 2002 Material
Each team was presented with 30 sets of student papers from the February 2002 administration, sequenced from low to high total holistic scores, with marks and pass/fail decisions shown. Each team member read all four responses in each set and determined whether the entire set deserved a "certain pass," a "reasonable pass," a "reasonable failure" or a "certain failure." Each team member did this independently and without discussion with other team members. The team members entered their decisions on data sheets with their names on them. Next, team members discussed their individual decisions and arrived at one group decision for each set. Each group also entered its consensus decision on the data sheet with its team name on it. This process was repeated with another 30 sets of student papers per team, though this time the papers were neither ordered nor presented with marks or pass/fail decisions.

Step 4: Examination of the October 2002 Material
The entire process used with the work from the February 2002 administration was repeated with that from the October 2002 administration. The results of all stages of the decision-making process were tabulated.

Step 5: Comparison of the February 2002 and October 2002 Material
Fifteen sets paired according to their total holistic scores were presented to the teams. The teams decided whether each set was a "certain pass," a "reasonable pass," a "reasonable fail" or a "certain fail." In addition, they indicated which of each pair was stronger. All decisions were tabulated.

Results for Writing
EQAO staff, in consultation with external testing experts, discussed the results of the standard-setting procedure. As mentioned previously, for each of the four writing tasks (summary, series of paragraphs expressing an opinion, news report, information paragraph) a score of 0, 1, 2, 3 or 4 was assigned using a holistic rubric. These scores were then given a weighting to reflect the equating process; the weightings were 0 = 0, 1 = 10, 2 = 20, 3 = 35 and 4 = 45. This differential weighting ensured that only students with consistent performances across the four writing tasks passed the writing component of the test.

A final passing score for the writing component was determined to be 100 out of a maximum of 180 points.

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, New Jersey: Educational Testing Service.

Assessment Systems Corporation. *XCALIBRE User's Manual.* http://www.assess.com.

Barnard, J. J. (September, 1996). In search for equity in educational measurement: Traditional versus modern equating methods. Paper presented at Association for the Study of Evaluation in Education in South Africa's national conference at the Human Sciences Research Council Conference Centre, Pretoria, South Africa.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Toronto: Holt, Rinehart and Winston, Inc.

Clustan. What is cluster analysis? Retrieved March 25, 2003, from the World Wide Web: http://www.clustan.com/what_is_cluster_analysis.html.

Education Quality and Accountability Office (2002). *Guide for Administering the Ontario Secondary School Literacy Test, October 2002.* Toronto, Ontario: Queen's Printer of Ontario.

— (2001). *Guide for Administering the Ontario Secondary School Literacy Test, October 2001.* Toronto, Ontario: Queen's Printer of Ontario.

Kolen, M. J. & Brennan, R. L. (1999). *Test Equating: Methods and Practices.* New York: Springer-Verlag.

Muraki, E., Himbo, C. M., & Lee,Y. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*(4), 325–337.

National Association of State Boards of Education (1999). Setting cut scores on large-scale assessment. *The State Board Connection: Issues in Brief, 19*(2), 1–12.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. H. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Sireci, S. G. (2001). Standard setting using cluster analysis. In G. J. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives* (pp. 339–369). Mahwah, NJ: Lawrence Erlbaum.

Smisko, A., Twing, J. S. & Denny, P. (2000). The Texas model for content and curricular validity. *Applied Measurement in Education, 13*(4), 333–342.

Traub, R. E. (1994). *Reliability for the Social Sciences: Theory and Applications*. Thousand Oaks, CA: Sage Publications.

Yang, W. & Houang, R. T. (1996). *The effect of anchor length and equating method on the accuracy of test equating: Comparisons of linear and IRT-based equating using an anchor item design.* Paper presented at the Annual Meeting of the American Research Association, New York, NY.